Sveučilište u Zagrebu
Filozofski fakultet
Ivana Lučića 3
10000 Zagreb

University of Zagreb
Faculty of Humanities
and Social Sciences
Ivana Lučića 3
10000 Zagreb, Croatia

Tel./Phone: +385 (0)1 6120-111; Fax: +385 (0)1 6156-879; www.ffzg.hr

Professor Marko Tadić
Department of Linguistics

Review of the PhD thesis *Rapid Development of Language Resources* by Marek Grác at the Faculty of Informatics, Masaryk University, Brno

The PhD thesis *Rapid Development of Language Resources* comprises 108 pages of which 4 pages include front matter and TOC, 11 pages list of 119 references and 14 pages cover two appendices. This leaves 79 pages for the thesis text which is divided in 6 sections (1. Introduction, 2. Introduction to language resources, 3. Methodology considerations, 4. Syntactic resource – BushBank, 5. Semantic resource – Sholva, 6. Conclusions and Future Work).

I am not acquainted with the exact procedure and best practices in writing PhD thesis at the Faculty of Informatics, so I will provide my opinion based on the practice we have at the Faculty of Humanities and Social Sciences, at the University of Zagreb.

Generally, the topic covered by the thesis is an interesting one and it would be of enormous help in rapid development of language resources that are lacking for many languages, particularly under-resourced ones. This language independent methodology can be applied to any language that as a prerequisite already has developed some tools for automatic processing at the level of morphology (PoS/MSD-tagging and/or lemmatisation). This is largely expected since the morpological language layer is one of the first ones attacked in development of language technologies for any language. In this respect, any new approach that can provide NLP researchers in certain language with quickly developed and annotated corpora, is more than welcome. The thesis describes the methodology and test it in two case studies, namely, CBB.desam10k and CBB.blog corpora. However, at least at the time of writing the thesis, CBB.blog corpus was not fully annotated (only 80% was completed and only by one annotator), so it didn't exhibit all the expected results of proposed methods. In the case of the resource Sholva, the prerequisites are not so frequently existing since its development is dependent on more developed resources such as vallency dictionary and/or wordnet for a given language.

The idea of specifying in annotation the least needed quantity of information certainly contributes to the idea of cheap and rapid development of language resources. However, as pointed out by author himself, this quantity of information is highly application dependent and in this way limits the usage of language resource for this or closely similar applications only. For other

applications introducing additional information in annotation could be less or more expensive and this we can't foresee in advance. Also, a very good idea is to mark somehow in annotation process the easy and difficult syntactic phenomena by a predefined mark.

The structure of the text of the thesis lacks some parts. Personally, I would like to see an introductory section that gives an overview of related work. In this way the references to related work are scattered around the text and no overview of previous research in this direction is given. Also, one of the key concepts – *syntactic bush* – is not introduced through a clear definition at the beginning of the relevant section, so it is left to reader to collect all its instances (term usages) and try to figure out the definition of this term/concept. It would be helpful if this and other key concepts were defined somewhere in the beginning of the thesis and then this definition obeyed throughout the whole thesis.

Some errors and weak methodological points in the very text of the thesis are marked in notes accompanying the PDF version, so I would like to draw the attention to some of them in the list below:

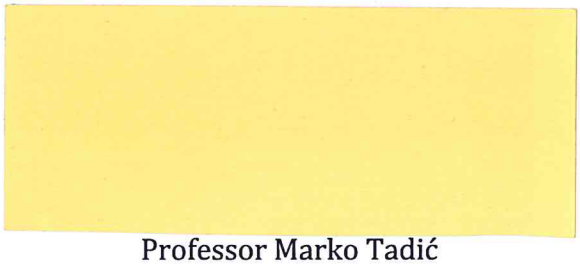
- Is there a reference for figures in the claim at page 16 "TEI attempts to create a standardized format which covers 90% of needs of 90% of users"?
- Could the claim at page 30 "Since we do not attempt to formalize intermediate steps, these syntactic trees tend to be more shallow than phrasal ones." be explained in more detail? What does it mean "more shallow" in this context since the syntactic analysis processes all tokens in the sentence, but in a different way. I would suggest to find a new term instead "shallow".
- Although at page 42 it was said that annotation of clause and coordination elements was fail, no explanation what happened with coordinations and clauses is given at the end. Any stats on these?
- Bad methodology at page 52 where a specialized grammar created by another researcher is taken, but no reference is given and no grammar is presented in appendices. In this way it is not possible to verify the research methods by a third party.
- Since CBB.blog is only 80% completed, how this fact reflects to the possible stats? Will it be possible to compare results from CBB.blog and CBB.desam10K at all? This is also a weak point in research performance of the methodology that author would like to demonstrate.
- What is missing in the whole section 4 is a table showing overall stats on CBB.desam10K and CBB.blog corpora, i.e. giving the figures on tokens, clauses, sentences, samples, etc. Additionally, these figures could be accompanied by information on number of tokens that are syntactically analysed, ambiguous etc. Also, few more real corpus examples with coordination, clauses detection, NP/PP and VP detection could be provided.
- Term usage at page 56: I don't want to go into a philosophical discussion, but where is the position of "word meaning" in the dichotomy between "word sense" and "word usage"? Are here "word meaning" and "word sense" synonyms, or we can find a place for additional "word meaning" as

different from "word sense"? See e.g. "Über Sinn und Bedeutung" by Frege.

- Types of evaluation at page 58: Here we can rise the queston of evaluation. Simple possibility of application of semantic tool to a morphologically and syntactically annotated corpus is not the proof of the quality of this semantic tool. Its application should be evaluated and this can be done intrinsically and extrinsically. The first approach would evaluate the performance of semantic tool alone, while the second should evaluate the quality of overall resource, i.e. how much the semantic tool contributes to the disambiguation of unsolved ambiguities at morphological and syntactic annotation level. So far, nothing is said about evaluation.
- Section 5.2.2, page 60: some important references on this topic are missing. For extracting relations between NP-VP-Attribute see several papers on extracting SVO-triplets by Marko Grobelnik and Dunja Mladenić. Also, papers by Paul Buitelaar on automatic ontology population from texts could be consulted here.
- Vague term "typical enough" at page 61: How do you define "typical enough", i.e. which criterion/criteria are used? Is it a prototypical meaning ("sense")? Is it a denotational meaning, or the most frequent meaning in which a word is used in a reference corpus? How about the fact that for some words metaphorical meaning is more frequent than their litteral meaning (this is mentioned later in thesis)?
- Missing reference [1] at page 83.

Having in mind these and other points marked in the PDF version of the text, that can certainly be corrected by the author, I believe this thesis could be accepted for obtaining the PhD degree at the Faculty of Informatics, Masaryk University in Brno.

With best regards,



Professor Marko Tadić
Head of the Chair of Algebraic
and Computational Linguistics